

PLENARY SPEECH

Data-driven learning: From *Collins Cobuild Dictionary* to ChatGPT

John Flowerdew^{1,2}

¹Lancaster University, Lancaster, UK and ²Birkbeck, University of London, London, UK
Email: johnflowerdew888@gmail.com

(Received 26 March 2024; accepted 4 April 2024)

Abstract

The invitation to speak at this conference on corpora and data-driven language learning (DDL) at COSEDI, the University of Grenoble, was received at the beginning of February 2023. ChatGPT had been released on 30 November 2022 and it was already becoming obvious that this technology would present revolutionary opportunities and challenges for corpus applications to language learning. Through a limited number of case studies – the *Collins Cobuild Dictionary*, a data-driven workshop for academic writing for research students, and the replication of the tasks used in the workshop, using ChatGPT – this presentation selectively highlights the trajectory of DDL from its beginnings to the present day and takes a look into a possible future with large language models (LLMs) such as ChatGPT.

1. Introduction

When I received this invitation to speak at this conference, I was very happy to accept because I like coming to France. I haven't been to France for many years, but my early childhood was spent in Jersey in the Channel Islands. So, I feel quite close to France.

Anyway, when I received the invitation back in February, I was very pleased to accept and the topic of data-driven learning (DDL) was fine for me. I was quite easily able to write the abstract that you have in your book of abstracts. But then, as time went by, I started to get a little bit concerned. I don't know if you can guess why? It wasn't because of the bedbugs we were reading about in the UK.¹ No, it was ChatGPT, which had come out a couple of months before and everyone was beginning to talk about. And it sounded like, 'wow, this is going to totally negate all the work that I've been doing for so many years with corpora. And what am I going to say at this conference?'

I had a conversation with my son one day asking him if he was using ChatGPT in his job. And he said, 'Oh, yeah, it's great. We do all our emails with ChatGPT.' And I told him about the problem I had – all that research that I've been doing that he'd had to listen to me talking about with my wife Lynne over the years. 'It's going to be redundant. I don't know what I'm going to say at the conference,' I said. He immediately replied, 'Well, it's easy. All you do is you talk about the beginning, and you show how we got to where we are now', which I thought was quite a good idea. So, I'm basically going to take that idea, but I'm going to simplify it a lot. I'm going to do three case studies. The first will be a very early example of corpus-based data-driven learning, John Sinclair and his work on the *Collins Cobuild Dictionary*, to show how indirect DDL started. Then I'll talk about a more recent study describing a workshop in DDL that I ran some years ago with a colleague

Revised version of a plenary presented at COSEDI (Specialised genre-based corpora): characterisation, methods, and pedagogical applications. University of Grenoble Alpes, France, 6–8 December, 2023.

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

(Chen & Flowerdew, 2018) where post-graduate students were introduced to corpus tools to help them with research writing. This will be an example of direct DDL. And then I'll talk about what would happen if you conducted that workshop again, but using large language models (LLMs), for example, ChatGPT.

I'll begin with a brief definition of data-driven learning (DDL), followed by the first case study and some more general comments about the contribution of John Sinclair. Next, I'll present the second case study describing the workshop. After that, after briefly introducing the notion of LLMs and their relation to corpus tools, I'll go on to case study 3, which will apply LLMs to the research writing workshop, which I will already have described. Finally, I'll talk in more general terms about the affordances and limitations of LLMs and end the talk with a brief conclusion.

2. What is DDL?

Very briefly, data-driven learning is the application of the affordances of corpus linguistics to language learning (see Boulton & Vyatkina, 2021; Dong et al., 2022 for recent research syntheses). This may apply in two ways, either indirectly as input to designing curriculum materials or directly as a learning tool where learners interact directly with corpus tools and corpora. I will cover both types in my case studies – indirect in case study 1 and direct in case studies 2 and 3.

3. Case study 1: John Sinclair and Collins Cobuild Dictionary

The first case study is the *Collins Cobuild English Language Dictionary*, published in 1987 and edited by John Sinclair (Sinclair, 1987). It's still going, now in its 10th edition, and called *Collins COBUILD Advanced Learner's Dictionary* (Collins Cobuild, 2023). Sinclair was a leader in our field. He had revolutionary ideas about corpora and language. Rather than summarise them myself, here are some comments from a relatively recent handbook article which talks about Sinclair's dictionary (Xu, 2022, p. 16). I separated the following as a series of quotations, but actually they are part of a single paragraph, but it is very dense and it tells us a lot about Sinclair and the *Collins Cobuild English Language Dictionary* and so it's good to dwell on each of the statements separately.

The Collins COBUILD English dictionary [CCD] is probably the bona fide game changer of dictionary making in the 20th century.

Corpus methodology is inherent in almost every bit of the dictionary.

For instance, the selection of head words is based on the frequency count of all English words in a 7.3-million-word corpus, initially called the main corpus, later referred to as Bank of English.

Actually, the Bank of English still exists. But it's a 450-million-word subset of the 4.5-billion-word COBUILD corpus of written and spoken English texts (Collins language lover's blog,). So, it's moved on a lot, but it was revolutionary at the time.

The main innovation of the CCD is its phraseological description of the entry word.

For example, the typical collocation of the word *brink*, namely, *on the brink of* is in the first place embedded in the whole sentence definition 'If you are *on the brink of* something, usually something important, terrible or exciting, you're just about to do it or experience it.'

This is important. Sinclair's corpus approach allows for whole-sentence dictionary definitions based on authentic language, as opposed to the fabricated examples of traditional dictionaries.

The contextualised definition is itself a condensed piece of learning material.

This is also very important. and why I cited the name *Advanced Learner's Dictionary* because from the very outset this dictionary was targeted at language learners, learners of English.

At the end of the entry, the colligational pattern is summarized as 'N-SING: usu. on/to/from the N of n'. The three prepositions separated by slashes are ordered according to their probability of occurrence in the corpus.

Two example sentences in the same dictionary entry, namely, 'Their economy is teetering *on the brink of* collapse' and 'Failure to communicate had brought the two nations *to the brink of* war', were taken from the Birmingham Corpus to illustrate the characteristic uses of *on the brink of* and *to the brink of*.

The co-occurrence of *brink* with *collapse* and *brink* with *war* implies the negative semantic prosody of the entry word.

Semantic prosody (and semantic preference) were key innovations of Sinclair in linguistic theory. And as far as language teaching is concerned, authentic chunks of language (based on corpus data) are essential for learners to get to grips with the nuances of meaning conveyed by this concept.

And finally, the extended unit of meaning model. That is the phraseological framework has been systematically implemented in the CCD.

Again, this is another important linguistic insight of Sinclair's. In fact, it underlies his whole theory of language, the idea that meaning is conveyed not just by words, but by lexical items, which are not just words, but may be a single word or a group of words (Stubbs, 2009).

One thing to say about this early corpus work is that Sinclair (as with the creators of LLMs) was a great believer in bigger is better. He wanted to have as much data as possible. But in those days, spoken data was very difficult to come by. And there were lots of restrictions on copyright of written data. So, you'll find in that early corpus a lot of the data was from novels which were out of copyright. So, people at the time were saying, 'well, you know, if you want to learn how to write a novel, you can use *Cobuild*.' But it didn't really give you much information about different genres. That's not the case now, because they've reorganised and redesigned the corpus, but that was the situation at the time.

So, just to summarise some of Sinclair's theoretical contributions more generally for us. His notion of the idiom principle – the idea that language is made up of more or less pre-fabricated units of meaning rather than just single open choices according to the Chomskyan view of language is very important (Erman & Warren, 2000). Sinclair had a notion that language is primarily made up of these prefabricated chunks, which is very important for us, and, of course, for LLMs.

Moreover, Sinclair's emphasis on how the collocations of a word affect its meaning – the 'extended unit of meaning' that was mentioned in that citation above – is very important. You don't need to just consider a single word to understand its meaning; you need to look at the words it typically occurs with. To comprehend a word, you need to consider the company it typically keeps, to paraphrase Sinclair's teacher, Firth (1957), and many other subsequent linguists.

Also very important are his notions of semantic preference and prosody. Semantic preference – how words are associated with particular other words in semantic fields and certain words more likely to occur together based on their shared meanings or associations within a given context. And semantic prosody – how certain words carry not only their literal meaning but also an additional evaluative or connotative sense based on their typical patterns of collocation, the inherent evaluative or emotional colouring that words may acquire based on their frequent co-occurrence with other words in specific contexts.

And finally, I'd also mention – although this is not perhaps a corpus linguistics notion so much, but for this conference where we're talking about genres is very important – one of the defining features of a genre is its staged nature; it has a schematic structure; it goes from one stage to the next. Sinclair's work with Malcolm Coulthard (Sinclair & Coulthard, 1975) is not talked about so much

nowadays, but their study of classroom discourse was revolutionary, because they analysed the structure of discourse above the level of the sentence. Their notion of exchange structure consisting of three moves – INITIATION, RESPONSE, FOLLOW UP – the typical three-part exchange in the classroom – was very important. Most genre-based work in our field nowadays cites Swales (1990) and his notion of the moves in the research article. I'll talk about this shortly. It's very significant.

It's interesting that Swales and Sinclair were both working in Birmingham at the same time. I often wondered whether Swales had taken the term *MOVE* from Sinclair. I did in fact ask him prior to this conference by email and he said he hadn't (Swales, personal communication 30 November 2023). But you know, they were both working at the same time in Birmingham. They were in different universities. But I think it was something in the air perhaps at the time. Certainly, the notion of larger units of discourse is very important and it can be dealt with in DDL, as we'll see later.

So, Sinclair's work on the *Collins Cobuild English Dictionary* was a groundbreaking application of data-driven learning, indirect data-driven learning.

4. More recent examples following Sinclair's groundbreaking work

I can give you some more recent DDL examples that apply Sinclair's ideas, so you can see his influence. The *Academic Phrasebank* (University of Manchester, n.d. (a)) is an open-source resource, available on- or off-line. It's a resource for academic writers and it gives you phraseological units that you can use in your research writing, organised according to the sections of the research article and broken down according to rhetorical moves.² It's basically organised according to the moves in Swales's research article introductions. A functionally organised text, the academic writer can use this to find relevant prefabricated chunks – extended units of meaning – if you like, to help with their writing.

Another example is the *Louvain EAP Dictionary* (LEAD) (Centre for English Corpus Linguistics, n.d.; Granger & Paquot, 2015). This is an online dictionary and, again, focuses on typical collocations. So, it's somewhat like the *Academic Phrasebank*, but is based on the British National Corpus (BNC) (British National Corpus, n.d.). It develops awareness of discipline-specific phraseology. You can select a particular discipline with this resource and it will give you language specific to that discipline. Again, it's based on the original ideas of Sinclair.

You are also probably all familiar with online mobile dictionaries and translation applications like *Reverso* (Reverso Technologies Inc., n.d.), *Linguee* (DeepL, 2023), among others. They quote concordance lines as well as the translations of words and phrases you look up. The examples look very much like those in the *Cobuild Dictionary*, based on key-word-in-context (kwic) concordance lines, except you get many more than in the dictionary. This allows the user to really get a feel for the various nuances in meaning. The data is based on large parallel corpora, translation memory, and/or data scraped from the internet. I use *Reverso* myself a lot for French and Spanish (Figure 1).

In my view, these apps are the most user-friendly direct applications of DDL, as they seamlessly provide concordance lines along with possible translations of the searched words or phrases, referred to as 'examples' (thereby getting away from some of the linguistic jargon that might be off-putting for learners). These apps use parallel corpora, but they also use various AI tools as well.

The important point to make about all of these texts and apps in this section is that they work on the same principles developed by Sinclair, in particular the idiom principle, the extended unit of meaning, and the key-word-in-context.

5. Case study 2: A DDL workshop in research writing for postgraduate students

This is a study that I conducted with Meilin Chen, who was originally my Ph.D. student and later my collaborator (Chen & Flowerdew, 2018). The authors describe how they introduced data-driven learning to Ph.D. students for research writing purposes in a territory-wide project in Hong Kong. The main purpose of case study 2 is not to discuss that workshop, but to prepare the way for case

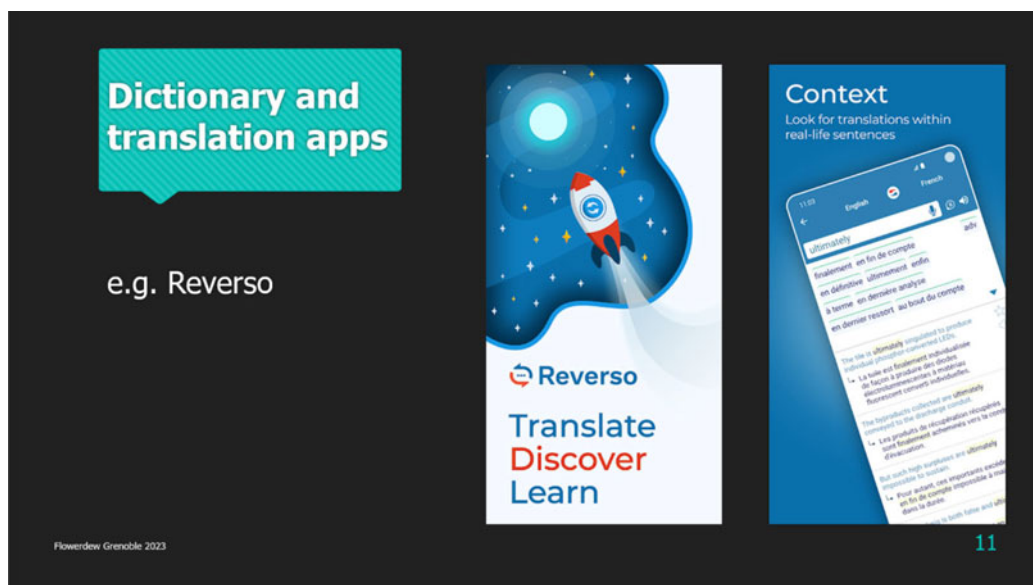


Figure 1. Reverso App showing kwic feature

study 3, which builds on case study 2. Secondly, I'll take the opportunity to demonstrate an application of another LLM, Microsoft Copilot (Microsoft, [n.d.](#)), which has just become available.

As I'm going to be talking about ChatGPT and other LLMs later in case study 3, I thought I would save myself summarising my own article and get Microsoft Copilot to do the job for me. Copilot is a chatbot based on ChatGPT. One of the functions of these chatbots is that they can summarise text. Microsoft Copilot suddenly appeared on my computer³ and so I thought I would use it. This is the summary of the article from Microsoft Copilot.

The project involved delivering 24 workshops on DDL to 473 Ph.D. students from six Hong Kong universities. The workshops aimed to introduce students to corpus tools and resources that can help them improve their research writing skills. The authors described the corpora and tools they used, the workshop outline, and the activities they implemented. They illustrated how they introduced students to BNCweb and AntConc, and how they guided students to explore lexico-grammatical and discoursal features of research writing using corpora. They summarised the students' evaluation of the workshop, which was highly positive and enthusiastic. They also discussed the students' suggestions for improvement and the challenges they faced in using corpora. They also explained how they provided follow-up support to students who wanted to continue practising their corpus skills.

Content-wise, this is quite an accurate summary, but it is missing one very important element of the workshop. As well as using BNCweb (BNCweb, [n.d.](#)) and AntConc, we also showed the students how to make their own corpus, their own tailor-made corpus, not mentioned by Microsoft Copilot. I think that's quite a serious criticism of the summarising function of Microsoft Copilot. Another minor point is that Microsoft Copilot didn't explain what BNCweb and AntConc are – rather important details for anyone not familiar with these applications. Just for the record, if you didn't already know it, BNCweb is the online version of the BNC and AntConc (Anthony, 2023) is Laurence Anthony's well-known concordancing programme.

As well as the content summary, the chatbot provided a synopsis of the feedback on the project.

The authors reported on the feedback they collected from the students through a questionnaire and interviews. They summarised the students' evaluation of the workshop, which was highly positive and enthusiastic. They also discussed the students' suggestions for improvement and the challenges they faced in using corpora.

The participants gave highly positive feedback on the workshops, rating them on seven statements and two open-ended questions. They found the workshops useful, friendly, high-quality, and informative. They also said they would recommend the workshops to others and use corpora in the future. The participants mentioned various aspects that they liked most, such as different corpus resources, useful expressions, hands-on activities, personal corpora, and writing conventions. They appreciated the affordances of corpora, the teacher's guidance, and the relevance to their writing problems.

The researchers interviewed three students who attended the workshops and found that they had different experiences and attitudes towards using corpora. One student was very enthusiastic and used BNCweb regularly. Another student was frustrated and switched to Google Scholar. The third student was selective and used BNCweb as a last resort. The cases showed that students' corpus use was influenced by factors such as time, computer skills, and availability of resources.

This again is quite an accurate summary. It's worth pausing here just briefly, because I think the three student cases referred to here are quite indicative of what happens when you try to persuade students to use corpora. Some of them are very enthusiastic, others not so enthusiastic. I think it's largely to do with how easy or how difficult it is to actually use the software.

A rather misleading aspect of this part of the summary is the statement that 'researchers interviewed three students who attended the workshops'. This strongly implies that only three students were interviewed, when in fact the three were selected from a larger sample of interviews as indicative of different attitudes. There are many reports online and in the literature of how LLMs tend to provide categorical statements that may be misleading and/or erroneous (Borji, 2023), sometimes referred to as HALLUCINATIONS. Certainly, this is an aspect of LLMs that unsuspecting students need to be aware of.

6. LLMs and corpus tools

L'éléphant dans la pièce is not a French expression. It's a literal translation of the English *The elephant in the room*. It refers to something you are very aware of, but you are not talking about for whatever reason. Interestingly, a common example translation in *Reverso* for 'elephant in the room' is '*éléphant dans un magasin de porcelaine*', which translates back into English as '*bull in a China shop*', which is something completely different. So, as an aside, this highlights a problem with dictionaries based on parallel corpora – a case of garbage in, garbage out. If the data in the translated corpus is not accurate, then the concordance examples based on it will also not be accurate. I've introduced this expression because LLMs are something that we need to be talking about in this meeting. And I think that if you're talking about LLMs, young people are going to know more about them than we as teachers do. So that's another *elephant in the room* that we need to consider as teachers in this era of Artificial Intelligence.

LLMs are applications that allow you to type in a question or command in ordinary language, referred to as a PROMPT. They produce an answer in ordinary language based on a huge training language database, or corpus in our jargon. It's an advance on Google and other search engines where key words produce a list of related websites. And it's a bit different to corpus linguistics, where a search word or group of words identifies all instances of that word or group in a corpus. LLMs seem to understand the question and provide an answer. Although they're using the same predictive principle as in Google – as Sinclair showed us so insightfully in linguistics – LLMs can participate in a whole conversation in this way. Bear in mind though that they never take the initiative – the perfect partner who never answers back, if you like. And they have no memory beyond the conversational thread they are participating in. Once you turn the thread off, it stops and restarts from the beginning, with no recollection of what has gone before.

7. Case study 3: Applying LLMs to the research writing workshop corpus queries

Let's see what happens if we use these LLMs to perform some of the tasks which we gave the students in those workshops that were summarised earlier; how we could use ChatGPT in that workshop if we wanted to do that.

In the workshop reported as Chen and Flowerdew (2018) and summarised here as case study 2 by Copilot, we began by brainstorming the following question with the students: *What verbs are used to report the results of a study?* Of course, nowadays, this could be done with ChatGPT. So, this question is also the first prompt that I gave ChatGPT in replicating the workshop. As Copilot only became

Prompt 1 ChatGPT:

- What verbs are used to report the results of a study?
- *When reporting the results of a study, various verbs and phrases can be used to convey the findings accurately. The choice of verbs often depends on the context, the type of study, and the tense you are using. Here is a list of common verbs and phrases used for reporting study results:*

Flowerdew Grenoble 2023

24

Verbs used to report the results of a study

1. Find: "The study **finds** that..."
2. Show: "The data **show** that..."
3. Indicate: "The results **indicate** that..."
4. Demonstrate: "The experiment **demonstrates** that..."
5. Reveal: "The research **reveals** that..."
6. Suggest: "The findings **suggest** that..."
7. Report: "The study **reports** that..."
8. Conclude: "The researchers **conclude** that..."
9. Confirm: "The results **confirm** that..."
10. Imply: "The data **imply** that..."
11. Establish: "The study **establishes** that..."
12. Illustrate: "The data **illustrate** that..."
13. Highlight: "The findings **highlight** that..."
14. Present: "The study **presents evidence** that..."
15. Support: "The results **support the hypothesis** that..."
16. Reflect: "The data **reflect** the relationship between..."
17. Corroborate: "The findings **corroborate** the earlier research."
18. Analyze: "The researchers **analyzed** the data and **found** that..."
19. Describe: "The study **describes** the pattern of..."
20. Compare: "The results are compared to previous studies, **showing**..."
21. Contrast: "In contrast to previous findings, this study **shows**..."
22. Examine: "The research **examines** the impact of..."
23. Investigate: "The study **investigates** the relationship between..."
24. Assess: "The findings **assess** the effectiveness of..."
25. Address: "The study **addresses** the question of..."
26. Explore: "The research **explores** the various factors affecting..."
27. Calculate: "The study **calculates** the statistical significance of..."
28. Measure: "The data **measure** the extent of..."
29. Discuss: "The researchers **discuss** the implications of the results."
30. Interpret: "The study **interprets** the data to mean that..."

Flowerdew Grenoble 2023

25

Figure 2. What verbs are used to report the results of a study? (ChatGPT)

available literally a few days ago, I didn't use it for this case study (except in a secondary role), instead using ChatGPT3.⁴ The two slides in Figure 2 represent the result from this prompt:

One interesting thing about ChatGPT that you can see from Figure 2 is that it gives you a little introductory paragraph, which actually I find really annoying, but you can ignore it. However, for our purposes, it's quite interesting to look at this anyway. When reporting, it tells you how to actually use the verbs. I didn't ask it to do this. I just asked it what verbs there were. But it gives you this preamble. I've actually colour-coded this somewhat and tidied it up. It doesn't look so neat and tidy in ChatGPT, but it's just to show you that it does give you a nice selection of verbs. The verbs it gives are as follows: *find, show, indicate, demonstrate, reveal, suggest, report, conclude*. This is very useful. It also gives you an example, or part of the sentence it's taken from, which I didn't ask it for. It seems to have a mind of its own. It seems to be saying: 'You probably want to know how to use these. So, I'll show you that as well.'

If you use Microsoft Copilot for the same prompt, the result is a bit different (Figure 3).

Again, it gives you the verbs, which I've highlighted. And it gives you examples, which I didn't ask it for. Now, one of the criticisms of LLMs is that you don't know what the data are that you get the answer from. But with Copilot, if you look at the bottom of the figure, it tells you where it's got the data (or where you can read up on the topic). So, interestingly, one of the sources is *Academic Phrasebank*, the resource I referred to earlier. So presumably, that book has been loaded into the training data. This highlights the fact that if you ask an LLM to tell you something, it's likely to tell you what it finds in a book or some other text. It doesn't actually explain it itself. So, for example, if you ask it for a grammatical rule, it just gives the answer from a grammar book; it doesn't actually create a rule itself (although if you manipulated the prompts enough it would probably be able to do that). So, artificial intelligence is not in fact INTELLIGENT at all, based on this data; it just regurgitates what is already in its database. (Either that or it decides that the internet will provide a better answer more easily than working out its own rule). LLMs are not able to truly understand concepts or make human-like inferences. Their responses are based purely on patterns learned from the data they were trained on (Borji, 2023).

So here in this example (Figure 3) it's doing the same thing with *Academic Phrasebank*. I actually contacted John Morley, the writer of this text, and asked him what he thought about it (personal

Responses from same prompt in Copilot

- some verbs that are used to report the results of a study are:
- **Shows**: This verb implies that the results provide clear and convincing evidence for a claim. For example, "The study **shows** that the new drug is more effective than the placebo." ¹
- **Indicates**: This verb suggests that the results point to a possible conclusion, but not as strongly as 'shows'. For example, "The data **indicate** that there is a correlation between sleep quality and academic performance." ²
- **Finds**: This verb reports the main outcome or discovery of the study. For example, "The research **finds** that meditation can reduce stress and anxiety." ³
- **Demonstrates**: This verb means that the results prove or illustrate something. For example, "The experiment **demonstrates** that the hypothesis is valid."
- **Reveals**: This verb means that the results disclose or uncover something new or surprising. For example, "The survey **reveals** that most students are satisfied with their courses."
- There are many other verbs that can be used to report the results of a study, depending on the context and the purpose of the report. You can refer to the links below for more examples and explanations of reporting verbs. I hope this helps! 😊 [Academic Phrasebank | Reporting results](#) ² [Reporting Verbs | Academic Skills Kit | Newcastle University](#) ³ [Academic Reporting Verbs - Academic English UK](#)

26

Figure 3. What verbs are used to report the results of a study? (Microsoft Copilot)

communication, 17 November 2023). He said he didn't know a lot about LLMs, but he was relaxed about his material being used, as long as it was helping people to write and that the creativity and sense of ownership that he sees with those who currently use it is not compromised. I thought that quite interesting. On the other hand, I went to a meeting in London recently with the Authors' Licensing and Collecting Society (ALCS), the society in the UK that collects royalties for writers, where there was a big debate about the ethics of writers' data being loaded onto these LLMs without permission or royalties paid. They were angry that their creative work was being used in this way. I think we will be hearing more about this issue in the coming weeks and months. There are already many court cases going on regarding this copyright issue.

Prompt 2: Give me examples of the verb *indicate* in research articles

- Certainly, here are some examples of how the verb "indicate" is used in research articles to report study findings:

Flowerdew Grenoble 2023

27

Examples of *indicate* in RAs

1. "The data collected from the survey **indicate** a strong correlation between regular exercise and reduced risk of cardiovascular diseases."
2. "The experiment results **indicate** that the temperature increase leads to a decrease in enzyme activity."
3. "The study's findings **indicate** a significant difference in test scores between the control group and the experimental group."
4. "The analysis of demographic data **indicates** a clear gender disparity in STEM professions."
5. "The research **indicates** a direct relationship between sleep quality and cognitive performance in young adults."
6. "The observations in our study **indicate** that the new drug treatment may have a positive effect on patients with a specific type of cancer."
7. "The data from the longitudinal study **indicate** that socioeconomic status is a predictor of educational attainment."
8. "The study's results **indicate** that there is a notable shift in consumer preferences towards sustainable and eco-friendly products."
9. "The survey responses from participants **indicate** a general dissatisfaction with the current healthcare system."
10. "The study's findings **indicate** that there is a need for further research to explore the underlying mechanisms behind these **observed** trends."

Flowerdew Grenoble 2023

28

Figure 4. Prompt 2: Give me examples of the verb *indicate* in research articles

As stated above, in the workshop we asked the students to brainstorm verbs to report results and then we showed them how they could search for these verbs using the corpus. Here are some examples of how the verb *indicate* is used if we ask ChatGPT to do this next task (Figure 4).

After the preamble (the first of the two slides in this figure), it gives us what looks like a concordance you might do with AntConc or some other concordancer: *indicate, indicate, indicate, indicates, indicates, indicate, etc.* . . . This is very nice. Of course, if you did this with a concordancer you'd have to use a wild card symbol – a symbol standing for any character – to get the different forms.

Now, prompt number three: *Which are the most frequent forms of the verb indicate in research articles?* In the workshop, we showed students how to search for that. You do that, of course, by getting frequency data based on the corpus you're using to compare the rate of occurrence of the different forms. ChatGPT gives you the different forms – *indicate, indicates, indicated, indicating* – but it gives you more information than you actually asked for again, telling you how these forms are used (Figure 5). On the other hand, it can't give you relative actual frequency data, because it's not working with a specific corpus.

The next task is to provide examples of the above forms of *indicate* (Figure 6), which it does accurately, with one example each for *indicate, indicates, and indicating*, and two examples of *indicated*. Perhaps, if the prompt had been worded better, it might have provided more examples for each form.

Moving on to the next prompt – *What words and phrases most often precede the verb indicate?* – we begin to see a bit of hallucinating (Figure 7). We see that as well as providing appropriate words – *results, findings, study, data, observations*, and so forth (highlighted), we also have examples in context, which is actually rather useful. However, three of the 11 examples provided (3, 6, and 7) are not with the verb *indicate*, but with other reporting verbs (*suggests, shown, and find*, respectively). Maybe it's been looking in *Academic Phrasebank* and/or other secondary sources and gone to the section with reporting verbs and found these examples. Furthermore, although the prompt asked for the most frequent verbs, there is no indication of relative frequency. This is again probably because it is not working on a dedicated corpus but is relying on secondary sources.

Prompt 3: Which are the most frequent forms of the verb *indicate* in research articles?

- In research articles, the verb "indicate" is often used in various forms to report study findings. The most frequent forms of the verb "indicate" include:
 1. "Indicate" itself.
 2. "Indicated" (past tense).
 3. "Indicates" (third person singular, present tense).
 4. "Indicating" (present participle/gerund form).
 5. "Indicated that" (used to introduce reported findings or results).
- These forms are commonly used to describe the findings, conclusions, or results of a study in a research article. The choice of form depends on the tense and structure of the sentence, as well as the context in which the verb is used.

Flowerdew Grenoble 2023 29

Figure 5. Prompt 3: Which are the most frequent forms of the verb *indicate* in research articles?

Prompt 4: Give me example sentences for the above forms of indicate

- Certainly! Here are example sentences for the different forms of the verb "indicate" in research articles:
 1. "The study's data **indicate** a strong relationship between smoking and lung cancer."
 2. "The research findings **indicated** a statistically significant difference between the control group and the experimental group."
 3. "The survey data **indicates** that a majority of respondents prefer online shopping over traditional retail."
 4. "The experiment's results, **indicating** a direct correlation, supported the hypothesis."
 5. "The analysis of the data **indicated** that there is a need for further investigation into this phenomenon."
- These examples demonstrate how "indicate" can be used in different forms to report findings and conclusions in research articles.

Flowerdew Grenoble 2023

30

Figure 6. Prompt 4: Give me example sentences for the above forms of *indicate*

Hallucinations are again apparent with the next task/prompt – *give me examples of indicate in research article conclusion sections* (Figure 8).

Again, it's given us some good examples with this prompt. You can see that these examples could be used in conclusion sections – *in conclusion, our study indicates that; to summarise, the results indicate; the findings of the research indicate* – but if you look at those that I've highlighted (5, 8, 9, 10) – *the data suggest; our investigation suggests; the study's findings provide valuable insights; this research*

Prompt 5: What words and phrases most often precede the verb *indicate*?

- In research articles, various words and phrases may precede the verb "indicate" to introduce findings, results, or conclusions. Common words and phrases that often precede "indicate" include:
 1. "The results **indicate** that..."
 2. "Our findings **indicate** that..."
 3. "The study **suggests** that..."
 4. "The data **indicate** that..."
 5. "These observations **indicate** that..."
 6. "As **shown** by the data,..."
 7. "In our analysis, **we find that**..."
 8. "According to the study, it **indicates** that..."
 9. "The study's conclusion **indicates** that..."
 10. "Based on the results, it is **indicated** that..."
 11. "Our research **indicates** that..."

Flowerdew Grenoble 2023

31

Figure 7. Prompt 5: What words and phrases most often precede the verb *indicate*?

Prompt 6: Give me examples of *indicate* in research article conclusion sections

1. "In conclusion, our study indicates that early intervention programs significantly improve developmental outcomes in children with autism."	6. "Our study's results indicate that a combination of lifestyle modifications and medication is the most effective approach to managing type 2 diabetes."
2. "To summarize, the results indicate that the new drug is a promising therapeutic option for managing chronic pain with minimal side effects."	7. "In summary, the research indicates that there is a significant gap in access to healthcare services in rural areas, necessitating policy changes."
3. "The findings of this research indicate a clear need for more targeted public health interventions to address the rising rates of obesity in our community."	8. "To wrap up, our investigation suggests that the use of a specific technology in manufacturing processes can lead to cost savings and increased productivity."
4. "In the final analysis, the study indicates that environmental factors play a crucial role in shaping the behavior of certain animal species."	9. "In the final assessment, the study's findings provide valuable insights into the potential therapeutic applications of herbal remedies for certain ailments."
5. "To conclude, the data suggest that the implemented educational reforms have had a positive impact on student performance and engagement."	10. "In conclusion, this research highlights the importance of social support networks in coping with stress and mental health challenges, as indicated by the data."

32

Figure 8. Prompt 6: Give me examples of *indicate* in research article conclusion sections

highlights the importance of – while these are all very useful also in conclusions, they're not with the verb *indicate*. So, again ChatGPT has fallen down here.

In summary, for case study 3, we've seen that LLMs (mainly ChatGPT, but also Microsoft Copilot) can perform all of the tasks selected from the workshop. The findings are impressive, but there are some misleading results. No doubt, more rigorous testing in future will give a much better evaluation. You might argue that it's not fair to evaluate LLMs on criteria for DDL. Some of the other things that they can do are incredible, including in the educational field. However, when commentators like Crosthwaite and Baisa (2023), whom I'll refer to shortly, are calling into question the whole future of DDL as a result of LLMs, it's a worthwhile exercise.

8. Affordances and limitations of LLMs

In this final section, let's talk a bit about the affordances and limitations of LLMs in more general terms, but still bearing DDL in mind. You may be interested in or already be familiar with these articles.

- Crosthwaite, P. & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast. *Applied Corpus Linguistics*, 3(3), 100066. <https://doi.org/10.1016/j.acorp.2023.100066>
- Lin, P. (2023). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, 3(3), 100065. <https://doi.org/10.1016/j.acorp.2023.100065>
- Kohnke, L., Moorhouse, B.L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal* 54(2), 537–550. <https://doi.org/10.1177/00336882231162868>

They've all appeared very recently, talking about GPT and data driven learning with corpora. All of them are very positive about how you might use ChatGPT, either directly in the classroom, as a direct application of DDL, or indirectly to design curriculum materials. I mean, ChatGPT can also design your lesson plans for you. Just search on the internet or use ChatGPT to find links showing you how to do it.

Specific forms of digital competence needed to use ChatGPT (Kohnke et al., 2023)

- Technological proficiency
 - Be aware of the features of ChatGPT
 - Understand how ChatGPT works
 - Construct effective prompts and interact with ChatGPT
 - Troubleshoot challenges using ChatGPT in the classroom
 - Stay up-to-date with changes to ChatGPT

Flowerdew Grenoble 2023

35

Figure 9. Specific forms of digital competence needed to use ChatGPT (Kohnke et al., 2023).⁵

Kohnke et al. (2023) talk about the technological competencies that students and teachers would need to use ChatGPT (Figure 9).

I think they were talking more about teachers, but these points can be applied to both teachers AND students. Kohnke et al. (2023) refer to ‘technological proficiency’ on this Figure. I think this a little misleading, because one of the great strengths of ChatGPT is that you can use ordinary language; you don’t really need any special technological proficiency. Under this heading, you need to BE AWARE OF WHAT CHATGPT CAN DO; you need to UNDERSTAND HOW IT WORKS; you need to be able to CONSTRUCT EFFECTIVE PROMPTS. I think that last point is crucial and there’s an important role for the

Ways to Improve Your English Using ChatGPT (based on Kohnke et al., 2023)

1) Using ChatGPT to practice English conversations

Example prompt: Can we engage in small talk, and you help correct my mistakes in English?

2) Using ChatGPT to create monologues and dialogues

Example prompt: Can you write a dialogue between four characters like a scene from the TV show [name of TV show]

3) Using ChatGPT to learn grammar

Example prompt: Can you write a short story about [topic] using the past tense and explain when to use the past tense in stories?

4) Using ChatGPT to learn vocabulary

Example prompt: Can you give me the most commonly used expressions in the business world?

5) Using ChatGPT to improve your writing

Example prompt: Can you rewrite this text [add text] more politely?

8) Use ChatGPT to prepare for a job interview

Example prompt: Can we have a back-and-forth interview for a primary teaching position?

Flowerdew Grenoble 2023

38

Figure 10. Ways to improve your English using ChatGPT (based on Kohnke et al. (2023))

teacher here. You need to TROUBLESHOOT ANY ISSUES THAT ARISE. When I showed you where ChatGPT gives you *suggest*, instead of *indicate* you're going to get that sort of thing turning up, referred to as, as I said, hallucination. Again, there's an important role for the teacher here. You need to STAY UP TO DATE WITH CHANGES. Yes, it's changing every day. What I'm telling you today is going to be out of date by tomorrow, if it's not already. None of these competencies are highly 'technological'; except perhaps the second one UNDERSTAND HOW CHATGPT WORKS, although even here, you only need a general understanding; they're just common sense or to do with general (first) language skills. However, I would emphasise that constructing effective prompts is the key to getting useful results (Lin, 2023).

Figure 10, which is also based on Kohnke et al. (2023), is a list of possible ways a student might use ChatGPT in their language learning.

You can see that these activities are not DDL as we know it. However, some of the suggested tasks could also be done with corpora. For example, number 4: [G]IVE ME THE MOST COMMONLY USED EXPRESSIONS IN THE BUSINESS WORLD. You could use the frequency function on a concordancer to do this if you had a large business corpus; or number 3: [E]XPLAIN WHEN TO USE THE PAST TENSE IN STORIES. You could certainly search for past tense forms in a corpus of stories; however, the learner would have to infer how to use them from the concordance lines provided rather than be presented with any 'rules'. But then you could also argue that such inductive tasks would have pedagogical benefits.

A number of generalisations can perhaps be made about the affordances and limitations of ChatGPT as a language learning tool, based on what these authors have to say. Figure 11 shows some of the advantages Crosthwaite and Baisa (2023) see for ChatGPT over corpus tools.

The first is the user experience. There is no doubt that ChatGPT is very easy to use, with millions of people already exploiting it daily with no specific training. Common criticisms of DDL are that the interfaces are difficult to use and that even after learners have been taught how to use them, they don't continue when they are left to their own devices (although see Charles (2014) on this second point). Another drawback with DDL is that the examples provided by corpora and presented as corpus lines may be difficult for learners with limited linguistic proficiency to understand (Lin, 2023). Given the right prompts, ChatGPT can tailor its output at an appropriate linguistic level. You can ask it to

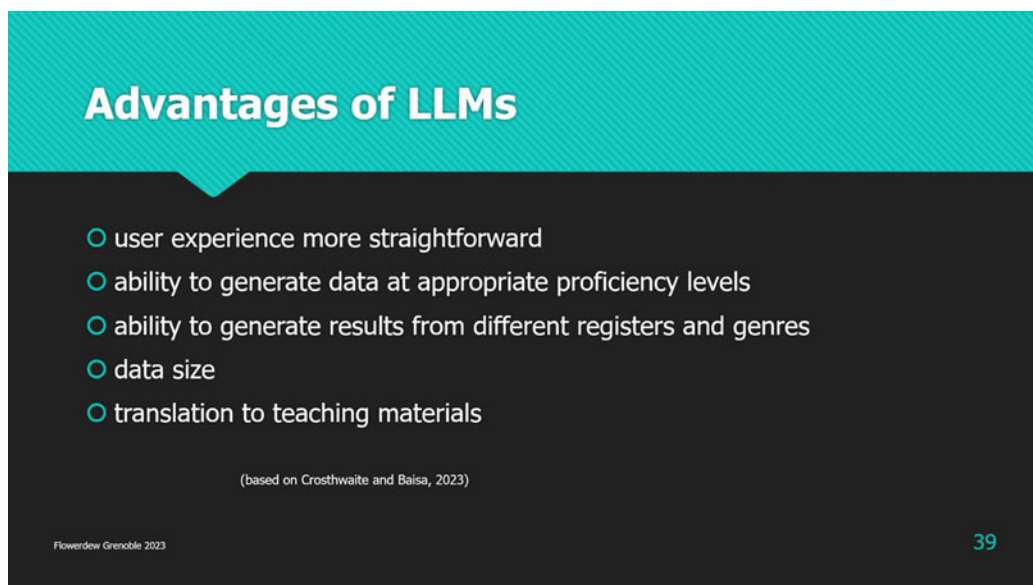


Figure 11. Advantages of ChatGPT (based on Crosthwaite & Baisa (2023))

Positive aspects of DDL over LLMs

1. Knowing the data
2. Authentic language
3. Replicability
4. Not the same copyright and other ethical issues that come with LLMs
5. Does not give false information (hallucinations)
6. (Language) learning and other cognitive and linguistic benefits of corpus consultation

(based on Crosthwaite and Baisa, 2023)

Flowerdew Grenoble 2023

47

Figure 12. Positive aspects of DDL (based on Crosthwaite & Baisa (2023))

write in simple English or more sophisticated English, American English, Indian English, or even Cockney, for example. Indeed, as a default, Copilot asks you to select a conversational style – MORE CREATIVE, MORE BALANCED, or MORE PRECISE. Furthermore, because they draw on huge databases, LLMs can produce output according to myriad styles – email, financial report, advertisement copy, and so forth (although some specialist genres that we are interested in with Languages for Specific Purposes (LSP) may not be represented (at least for the time being). Finally, Crosthwaite and Baisa (2023) cite the ability to create teaching materials as an advantage of LLMs, although this is not directly an issue for DDL.

In spite of these advantages of LLMs, Crosthwaite and Baisa (2023) still see a role for DDL as we know it, as shown in Figure 12.

Crosthwaite and Baisa (2023) argue that an advantage DDL has over LLMs is that with the former you know the data that you are working with. As already noted, the sources of the outputs produced by LLMs are unknown (or we only make guesses as to what they might be), in contrast to DDL, where we can be confident that the language that we are asking learners to work with is appropriate to our goals. Learners are also able to consult the broader context of the corpus lines they are perusing, if they so wish, by clicking on the key word in context. Furthermore, learners may build their own personalised corpora, as illustrated in Chen and Flowerdew (2018) (see also Lee & Swales, 2006).

Related to knowing the data, users of corpora can be sure that the language they are working with is authentic. Authentic language, indeed, was one of the original arguments in favour of using corpus-derived data in language teaching and learning in the first place (Gilmore, 2007). As already noted, we are not sure where the data comes from with LLMs and, unlike with corpora, the language generated with LLMs does not correspond to actual attested use but is created according to statistical relationships. Related to this point also, corpus queries are replicable. We can do the same search and come up with the same findings, unlike with LLMs, which give a different answer even if you give it the same query, something that can be confusing to learners. Furthermore, corpora do not make mistakes, whereas LLMs are liable to hallucinate and give false information, such as when in our third case study ChatGPT gave us examples of the use of the verb *indicate* in conclusion sections that did not include the word *indicate*! Of course, with corpora, THE USER can make mistakes (in choosing the corpus and tool, in formulating queries, or interpreting the results) and the corpus is only as good as what it contains. But, the corpus itself does not give false information.

Finally – and this is the point that, along with Crosthwaite and Baisa (2023), I want to emphasise – the procedures learners perform in DDL promote language learning in a way that LLM prompts do not. In DDL, learners discover how to search for patterns and induce meanings from many instances of use. These patterns of form and meaning can be added to the learner's linguistic repertoire. As various authors have argued, by analysing recurrent patterns, learners apply inductive processes involving noticing, language awareness, hypothesis-making and -testing, sensitivity to linguistic variation, and various data skills (e.g., Flowerdew, 2015; O'Keeffe, 2021; Pérez-Paredes et al., 2019; O'Sullivan, 2007). With LLMs, in contrast, users are provided with answers to queries, but there is no inductive learning. It is a transmission model, with the learner asking the question and the LLM providing the information. There is little cognitive work needed on the part of the learner.⁶

9. Conclusion

In this talk, I've presented three case studies – an early ground-breaking indirect application of DDL with the *Collins Cobuild Dictionary* (along with some other examples using the same insights employed in *Cobuild*); a more recent direct application, with the workshop I conducted with Meilin Chen; and a replication of some of the queries from that workshop but using LLMs (mainly ChaptGPT) – and made some generalisations about DDL and LLMs. One wonders what Sinclair might have thought of ChatGPT if he had lived to see its development. As I previously mentioned, he favoured large corpora, and ChatGPT certainly uses a very large corpus indeed. Sinclair required a large corpus, however, because he wanted to corroborate linguistic patterns and meanings that he found by means of large numbers of examples. In this way, he achieved deterministic results. Having found many examples in his (large) corpus, he could be confident that he would likely find the same patterns and meanings in other linguistic data. LLMs are not deterministic like this, however; they are probabilistic rather than deterministic. You can ask them the same question many times and each time they will give you a different answer. Furthermore, as we have seen, they may provide you with false data and linguistic examples that have been made up. It is not surprising, therefore, that in his memorial paper on the contribution of John Sinclair, Stubbs (2009) tells us that Sinclair was dubious about Natural Language Processing (the precursor to LLMs).

I don't think, however, that Sinclair could have predicted the paradigm shift that we are experiencing today with LLMs. There is still a place for DDL as we know it, especially in indirect approaches, where deterministic results from linguistic analysis feed into learning materials, as with the *Cobuild* dictionary, but also in classroom concordancing, where, along with Tim Johns (1997), we want every learner to be a Sherlock Holmes and (guided by the teacher) find out for themselves how the target language works and add that knowledge to their linguistic repertoire.

The answer lies in a synthesis of both approaches. In the article I cited by Lin (2023), she describes how she has experimented to produce prompts for ChatGPT to make it act like a concordance (although the prompts she had to produce through trial and error were complicated and their production extremely time-consuming). 'I was particularly interested in exploring the feasibility of using ChatGPT as a concordancer,' she states, 'because this might allow learners and teachers to benefit from both concordancing and ChatGPT' (p. 5). Similarly, Laurence Anthony told me in an email yesterday that he is introducing a function into his AntConc app that allows the use of ChatGPT alongside the usual concordancer functions. This is surely the way to go. Future research and practice will be required, of course, however, to show how such a synthesis can come about and what it would look like. How does one know when to use one or the other? What are the advantages of each (in terms of producing textual output and in terms of language acquisition), and how do they relate to theories of language learning? Laurence also told me something that will make a good quote on which to close this talk:

Generally speaking, I think most people in linguistics don't really understand how LLMs work and what their potential is. So much about LLMs is a mystery, but they are incredibly powerful and will change everybody's lives. As linguists, it seems reasonable to me that we should be at

least somewhere near the front line when it comes to understanding how these large LANGUAGE models are working, proposing new ways to use them, and improving them for the future. (Laurence Anthony, personal communication, 4 December 2023)

So, there you have it.

Declaration of use of generative AI and AI-assisted technologies in the preparation of this article

During the preparation of this work, the author used Microsoft Copilot in the production of the summary used in the second case study and ChatGPT and Microsoft Copilot for the production of the data used in the third case study. He also used ChatGPT and Copilot for literature search. He furthermore tried the two apps for reference production, but they were very inaccurate.

Acknowledgment. The author gratefully acknowledges discussion about LLMs with Laurence Anthony, Peter Crosthwaite, and Richard Forest prior to making this presentation. The author takes full responsibility for its content. He also gratefully acknowledges feedback from three reviewers and the editor of *Language Teaching* in revising the speech for publication.

Notes

- ¹ Several months prior to the conference, there had been widespread reports of outbreaks of bedbugs in France, including in Grenoble, the conference venue.
- ² The corpus used for this resource is described as follows: ‘The original corpus from which the phrases were “harvested” consisted of 100 postgraduate dissertations completed at the University of Manchester. However, phrases from academic articles drawn from a broad spectrum of disciplines have also been, and continue to be, incorporated.’ (University of Manchester, (n.d.) (b)).
- ³ Microsoft Copilot is a Microsoft application of GPT4. In January 2024, Microsoft made available a dedicated Copilot key on Windows 11 keyboards. It is also available to subscribers to Microsoft 365, where the author of this speech accessed it.
- ⁴ Both ChatGPT and Copilot are based on Open-AI’s GPT LLM. As mentioned, the version of ChatGPT used in the case study was ChatGPT3, based on the GPT3 LLM, while Copilot uses GPT4, so is more sophisticated. Since the conference, GPT4 has become generally available in the (subscription-based) ChatGPT4 chatbot.
- ⁵ The points in this table are part of a larger table in Kohnke et al. (2023, Table 1, p. 546).
- ⁶ That is not to say that, given the right prompts, an LLM could not mimic a concordancer (see Lin, 2023), but it would be a very complex procedure.

References

- Anthony, L. (2023). *Antconc (version 4.2.4) [computer software]*. Waseda University. Retrieved January 1, 2024, from <https://www.laurenceanthony.net/software>
- BNCweb. (n.d.) Retrieved January 1, 2024, from <http://bncweb.lancs.ac.uk/cgi-inbncXML/BNCQuery.pl?theQuery=search&urlTest=yes>
- Borji, A. (2023). A categorical archive of ChatGPT failures. *Information & Culture: A Journal of History*, 47(4), 414–456. <https://arxiv.org/abs/2302.03494>
- Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions. *Language Learning and Technology*, 25(3), 66–89. <http://hdl.handle.net/10125/73450>
- British National Corpus. (n.d.) Retrieved January 1, 2024, from <http://www.natcorp.ox.ac.uk/>
- Centre for English Corpus Linguistics. (n.d.). Retrieved January 1, 2024, from LEAD | Corpus search (<https://corpora.uclouvain.be/cecl/lead/#?genre=anthro>)
- Charles, M. (2014). Getting the corpus habit: EAP students’ long-term use of personal corpora. *English for Specific Purposes*, 35, 30–40. doi:10.1016/j.esp.2013.11.004
- Chen, M., & Flowerdew, J. (2018). Introducing data-driven learning to PhD students for research writing purposes: A territory-wide project in Hong Kong. *English for Specific Purposes*, 50, 97–112. doi:10.1016/j.esp.2017.11.004
- Collins Cobuild. (2023). *Collins COBUILD advanced learner’s dictionary* (10th ed.). Collins. <https://collins.co.uk/products/9780008444907>
- Collins language lover’s blog. “The history of COBUILD.” Collins. Accessed January 4, 2024. <https://blog.collinsdictionary.com/the-history-of-cobuild/>
- Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast. *Applied Corpus Linguistics*, 3(3), 100066. doi:10.1016/j.acorp.2023.100066
- DeepL, S. E. (2023). *Dictionary Linguee (Version 1.3.0) [Mobile application software]*. Retrieved January 10, 2024, from https://play.google.com/store/apps/details?id=com.linguee.linguee&hl=en_GB&gl=US

- Dong, J., Zhao, Y., & Buckingham, L. (2022). Charting the landscape of data-driven learning using a bibliometric analysis. *ReCALL*, 35(3), 1–17. doi:10.1017/S0958344022000222.
- Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text & Talk*, 20(1), 29–62. doi:10.1515/text.1.2000.20.1.29
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in linguistic analysis* (pp. 1–31). Special Volume of the Philological Society.
- Flowerdew, J. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska, & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15–36). John Benjamins.
- Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching*, 40(2), 97–118. doi:10.1017/S0261444807004144
- Granger, S., & Paquot, M. (2015). Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica*, 31(1), 118–141. doi:10.1515/lexi-2015-0007
- Johns, T. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). Longman.
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, 54(2), 537–550. doi:10.1177/00336882231162868
- Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25(1), 56–75. doi:10.1016/j.esp.2005.02.010
- Lin, P. (2023). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, 3(3), 100065. doi:10.1016/j.acorp.2023.100065
- Microsoft. (n.d.). Copilot. Microsoft Edge. Retrieved January 1, 2024, from <https://www.microsoft.com/en-gb/edge/features/copilot?form=MA13FJ>
- O’Keeffe, A. (2021). Data-driven learning – a call for a broader research gaze. *Language Teaching*, 54(2), 259–272. doi:10.1017/S0261444820000245
- O’Sullivan, Í. (2007). Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL*, 19(3), 269–286. doi:10.1017/S095834400700033X
- Pérez-Paredes, P., Sánchez-Tornel, M., & Aguado-Jiménez, P. (2019). Data-driven learning: A pedagogical model for language learning and teaching. In P. M. Sullivan, T. L. Lantz, & B. A. Sullivan (Eds.), *Handbook of research on integrating technology into contemporary language learning and teaching* (pp. 1–22). IGI Global. doi:10.4018/978-1-5225-7825-1.ch001
- Reverso Technologies Inc. (n.d.). Reverso [Version 9.9.7]. Retrieved January 1, 2024, from <https://context.reverso.net/translation/>
- Sinclair, J. (1987). *Collins COBUILD English language dictionary*. Collins.
- Sinclair, J. McH. & Coulthard, R. M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford University Press.
- Stubbs, M. (2009). Memorial article: John Sinclair (1933–2007): The search for units of meaning: Sinclair on empirical semantics. *Applied Linguistics*, 30(1), 115–137. doi:10.1093/applin/amn052
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.
- University of Manchester. (n.d.) (a). Academic phrasebank. Retrieved January 1, 2024, from <http://www.phrasebank.manchester.ac.uk/>
- University of Manchester. (n.d.) (b). Academic phrasebank. Retrieved January 1, 2024, from <https://www.phrasebank.manchester.ac.uk/about-academic-phrasebank/>
- Xu, J. (2022). A historical overview of using corpora in English language teaching. In R. R. Jablonkai, & E. Csomay (Eds.), *The Routledge handbook of corpora and English language teaching and learning* (pp. 11–25). Routledge. doi:10.4324/9781003002901

John Flowerdew is a visiting professor at Lancaster University and an honorary research fellow at Birkbeck, University of London. Previously he was a professor at City University of Hong Kong and at University of Leeds. Notable more recent books are *Introducing English for research publication purposes* (with P. Habibie) (Routledge); *Signalling nouns in discourse: A corpus-based discourse approach* (with R.W. Forest) (CUP); *The Routledge handbook of critical discourse studies* (with J. Richardson); and *Discourse in English language education* (Routledge). He is active in research and publication and is regularly invited to speak at international conferences.